LINDAT/CLARIAH-CZ

PAVEL STRAŇÁK

Repository

lindat.cz/repository

LINDAT/CLARIAH-CZ

PAVEL STRAŇÁK

FAIR repository for language data

+ humanities' datasets

lindat.cz/repository

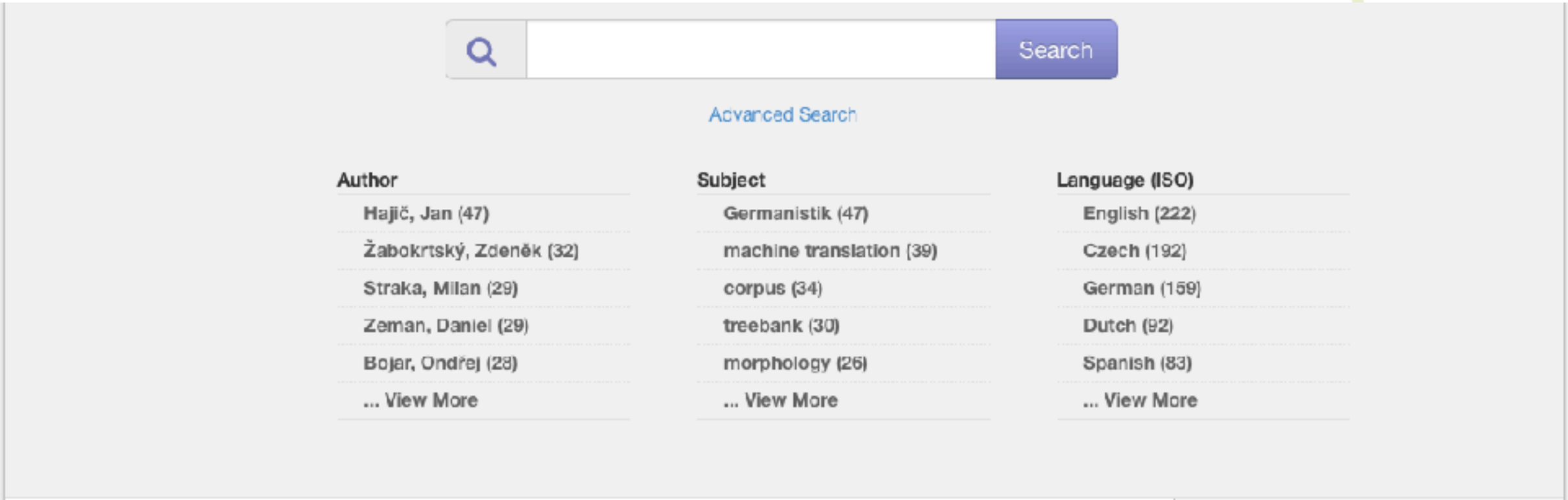
CLARIN

CLARIN-DSpace GitHub project

- https://github.com/ufal/clarin-dspace
- About 15 installations in 11 countries
  - Some installations just using the Github documentation and #dspace at CLARIN Slack
  - Some installations by LINDAT team using CLARIN Mobility grants. With some preparation in advance, a working B-centre compliant repository can be up and running in 2 days.
- Some forks (Polish, Norwegian) have a few additional features
- Most centres probably motivated by taking a ready-made repository
  - The collaborative development (naturally) works a bit less great than we had hoped
  - Upcoming upgrade to Dspace 7 (new UI) will be the real test of the community

Data Repository

Preserve and find language data and NLP tools



What's New

ToolService

LINDAT / CLARIN

CorpusExplorer

Author(s):
Rüdiger, Jan Oliver

What can you do?

DEPOSIT

CITE

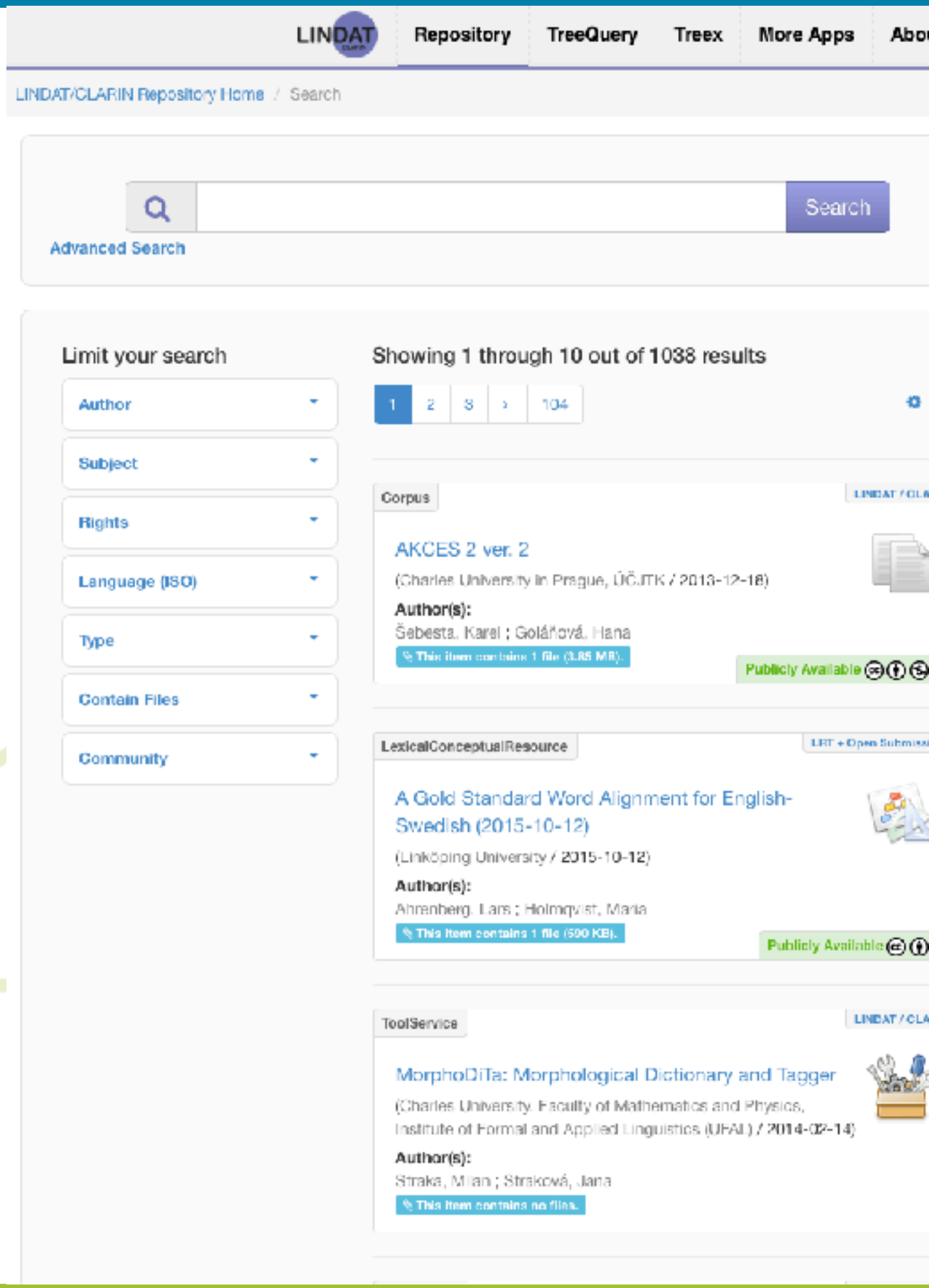
Browse

Data Repository

- OPEN ACCESS (as can be – Public License Selector)
- > 500 registered users
  - submitters & users signing licenses (not everything can be OA)
- 200+ Data Records
  - > 1000 Metadata Records
  - 80 languages
- 100 TB+ Data in Repository (+ 1PB of UCS Shoah Foundation Archive)

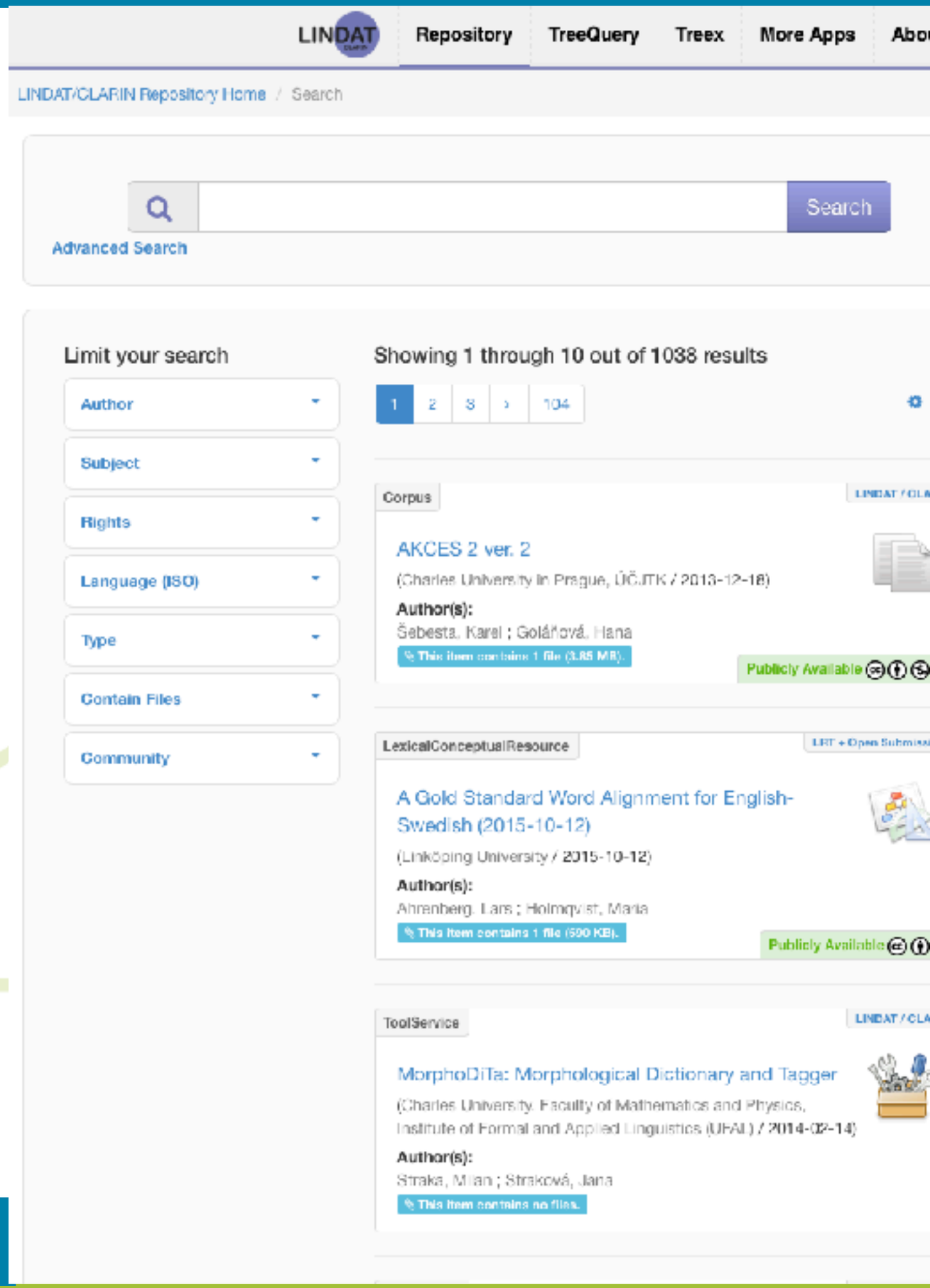
Data Repository

- > 500 registered users
  - submitters & users signing licenses (not everything can be Open Access)
- 200+ Data Records
  - > 1000 Metadata Records
  - 80 languages
- 100 TB+ Data in Repository (+ 1PB of UCS Shoah Foundation Archive)



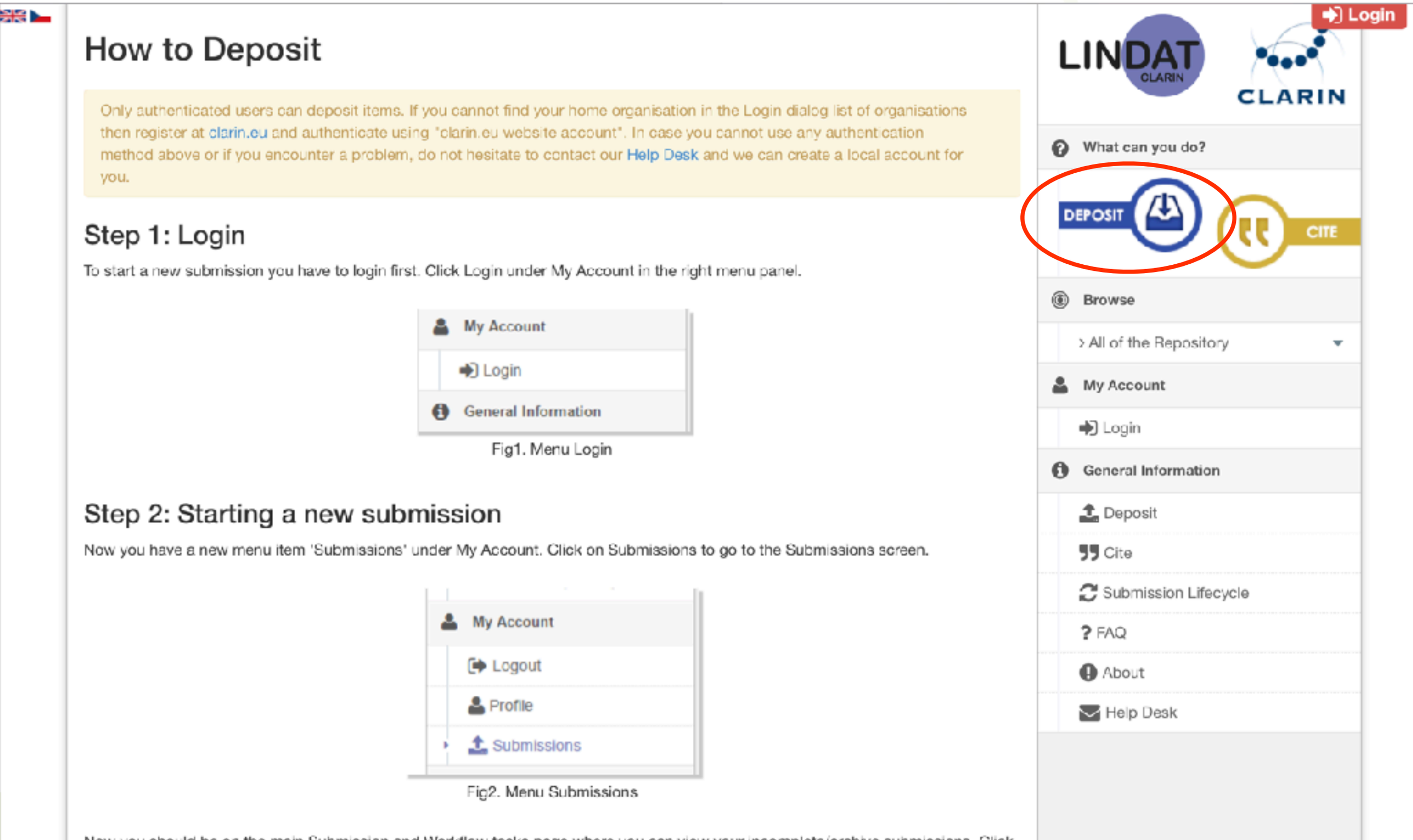
Data Repository

- Safe preservation
  (upload and don't worry)
- Discovery & Reuse
- Direct data citation
  (works in Google Scholar)
- Licensing
  (Open Access, but also more options)
- Versioning
- Worldwide (for everyone), easy to use



Safe Preservation

Upload and don’t worry



Deposition Guide

step-by-step description

1. login
2. fill-in metadata
3. drag&drop data
4. select a license
5. submit

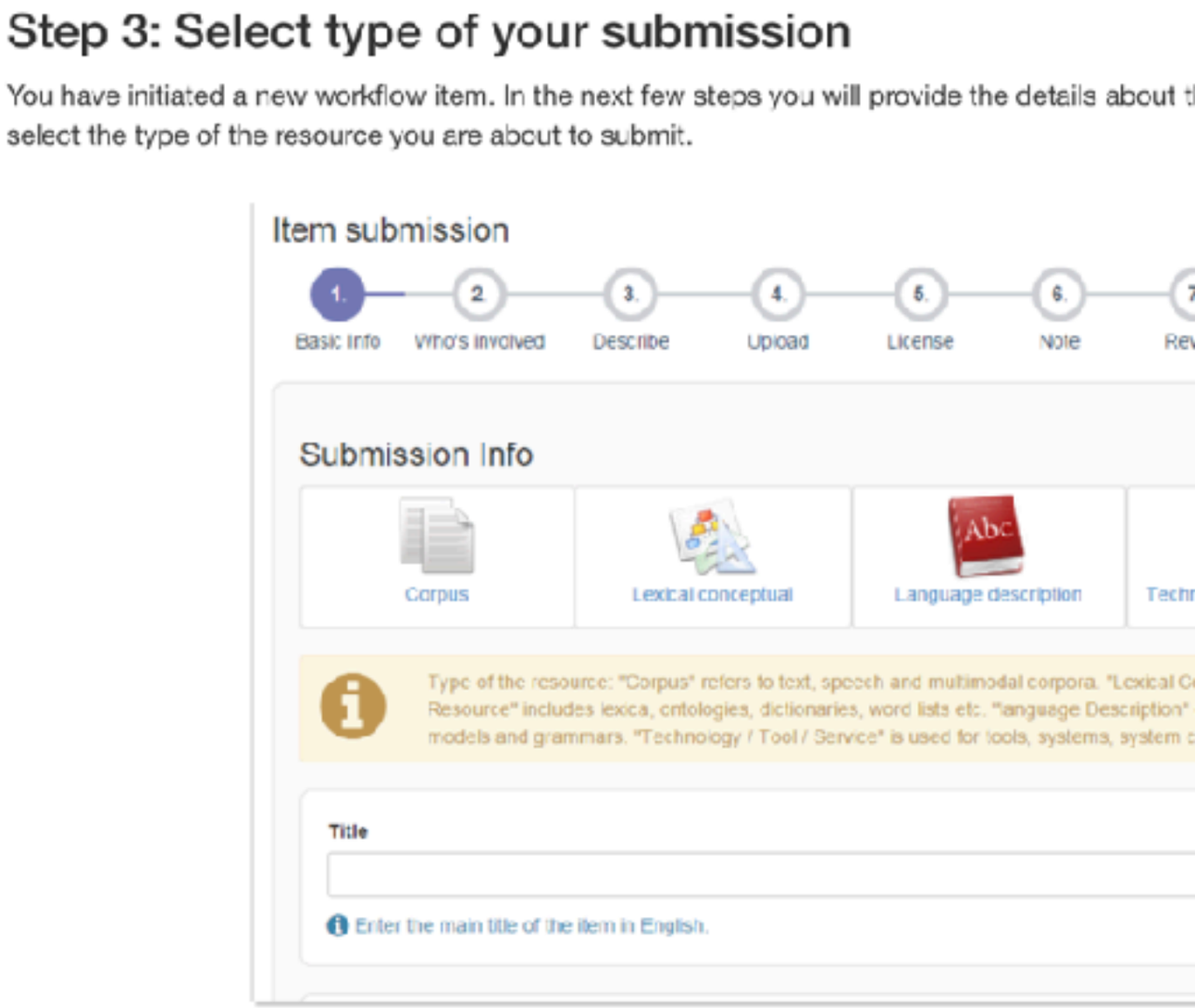
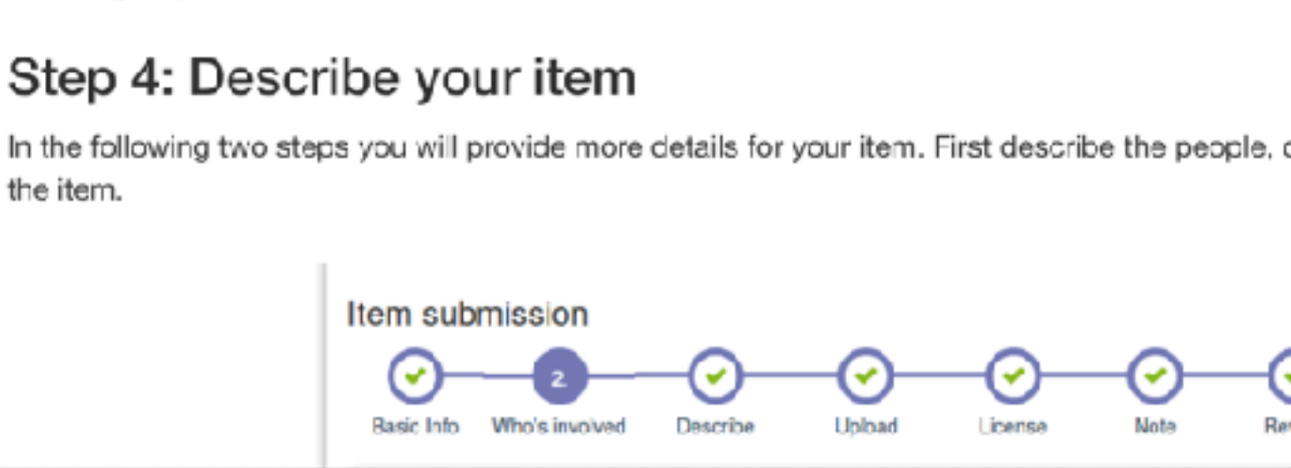


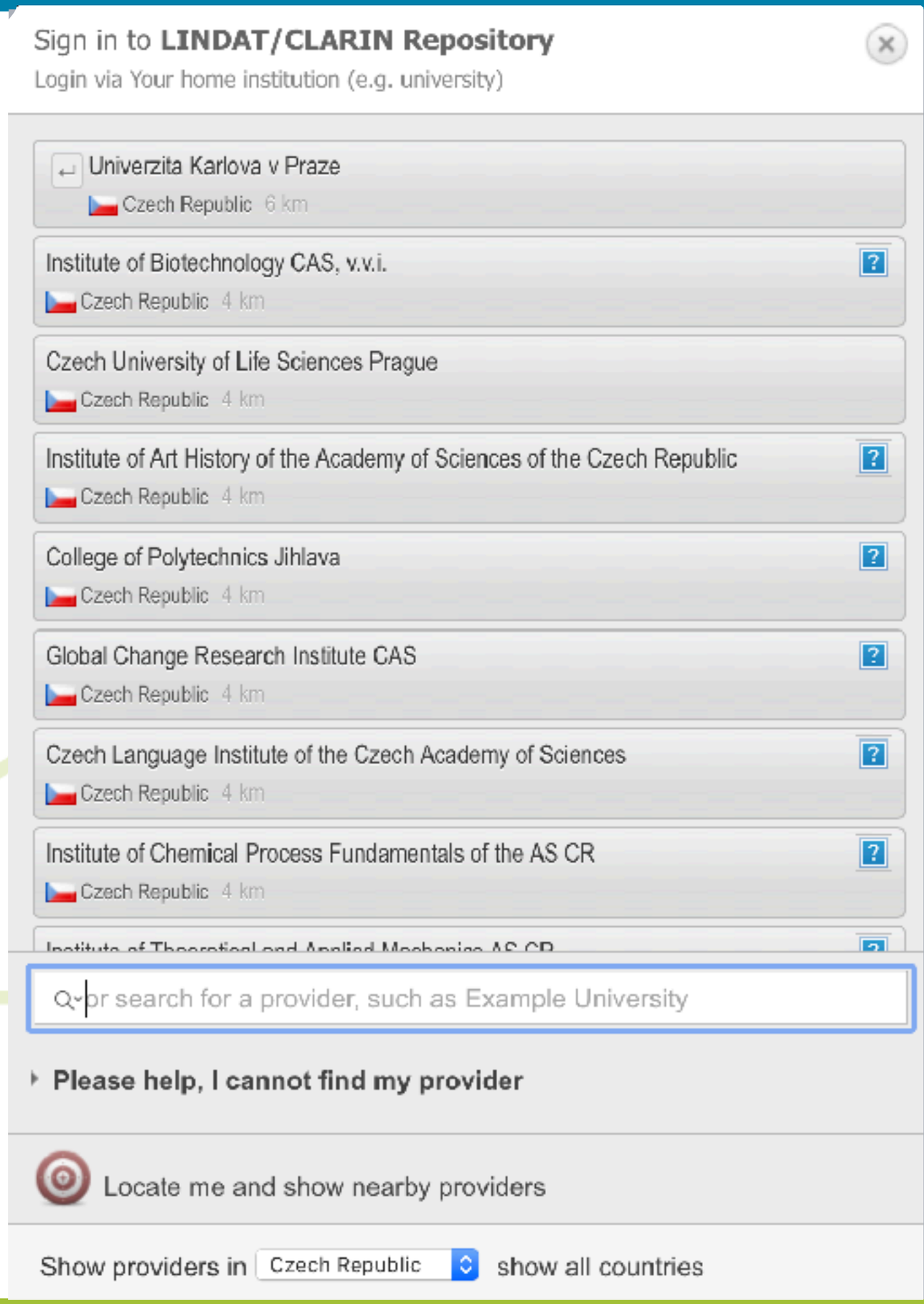
Fig4. Submission info

Click on one of the type buttons e.g. Corpus. Proceed with filling the basic information such as t
following step.

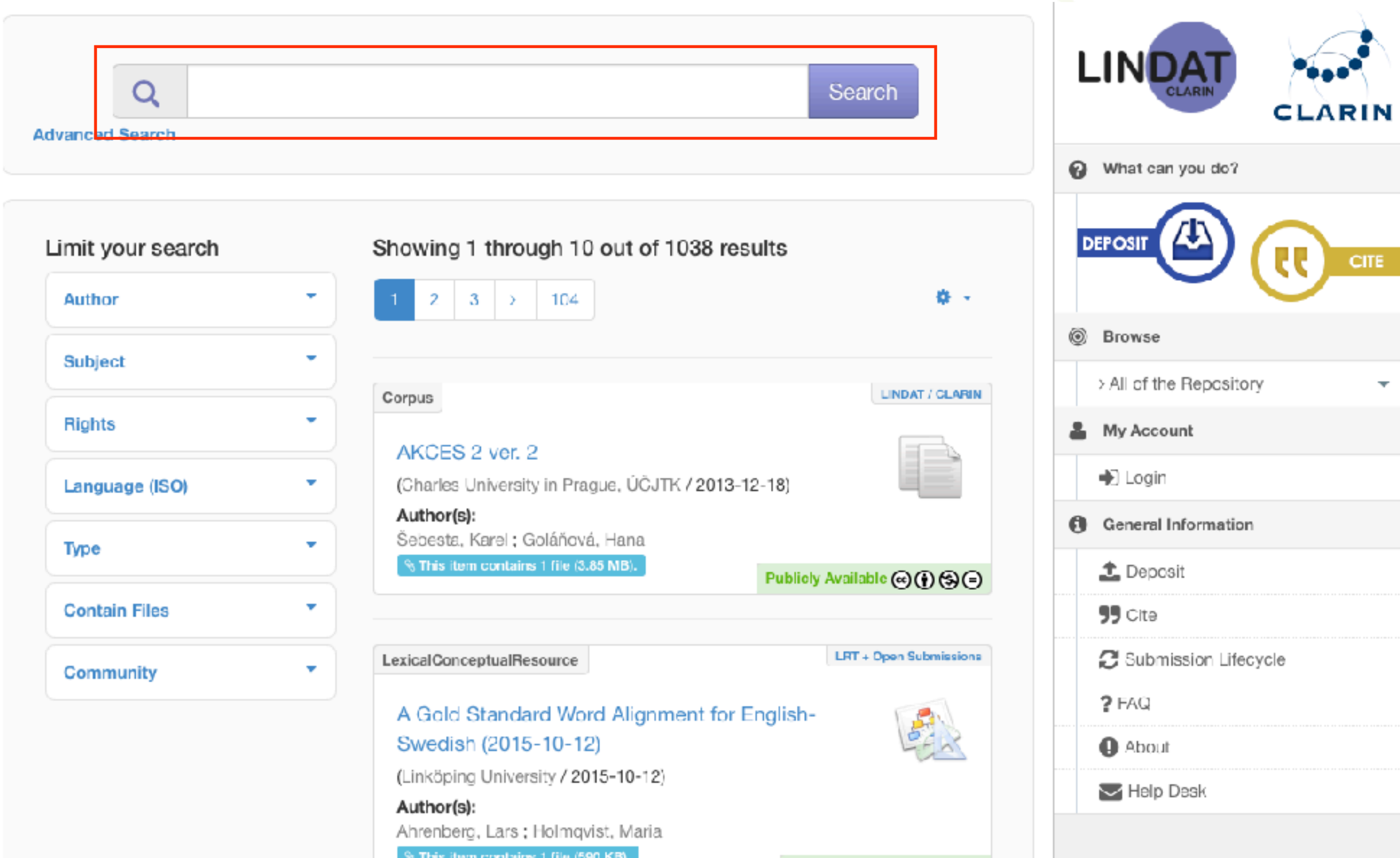


Login to Deposit

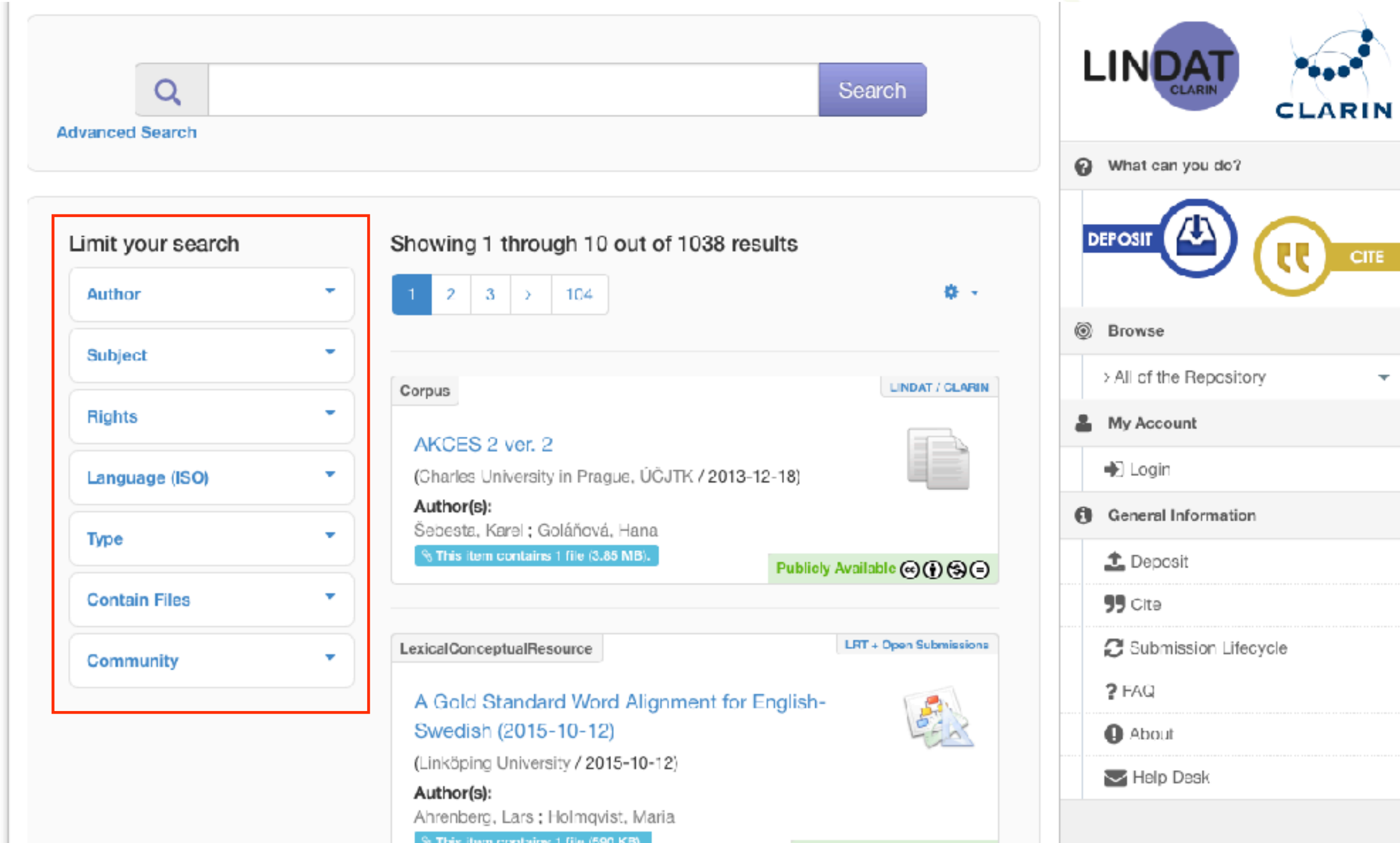
- institutional logins (EduID-cz, EduGAIN)
- CLARIN account for the “homeless researchers”
- minimal personal info



Faceted Search

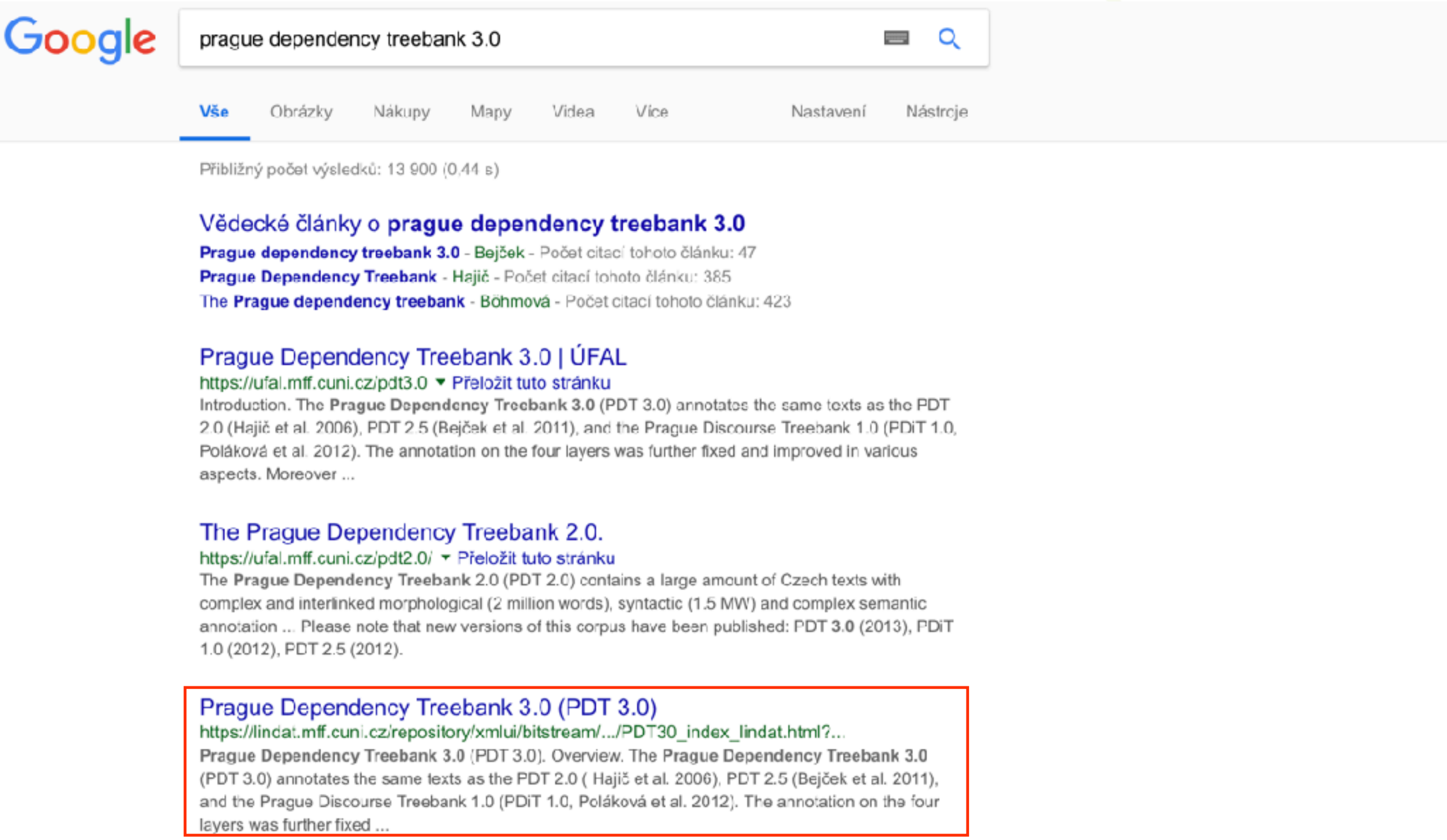


Faceted Search



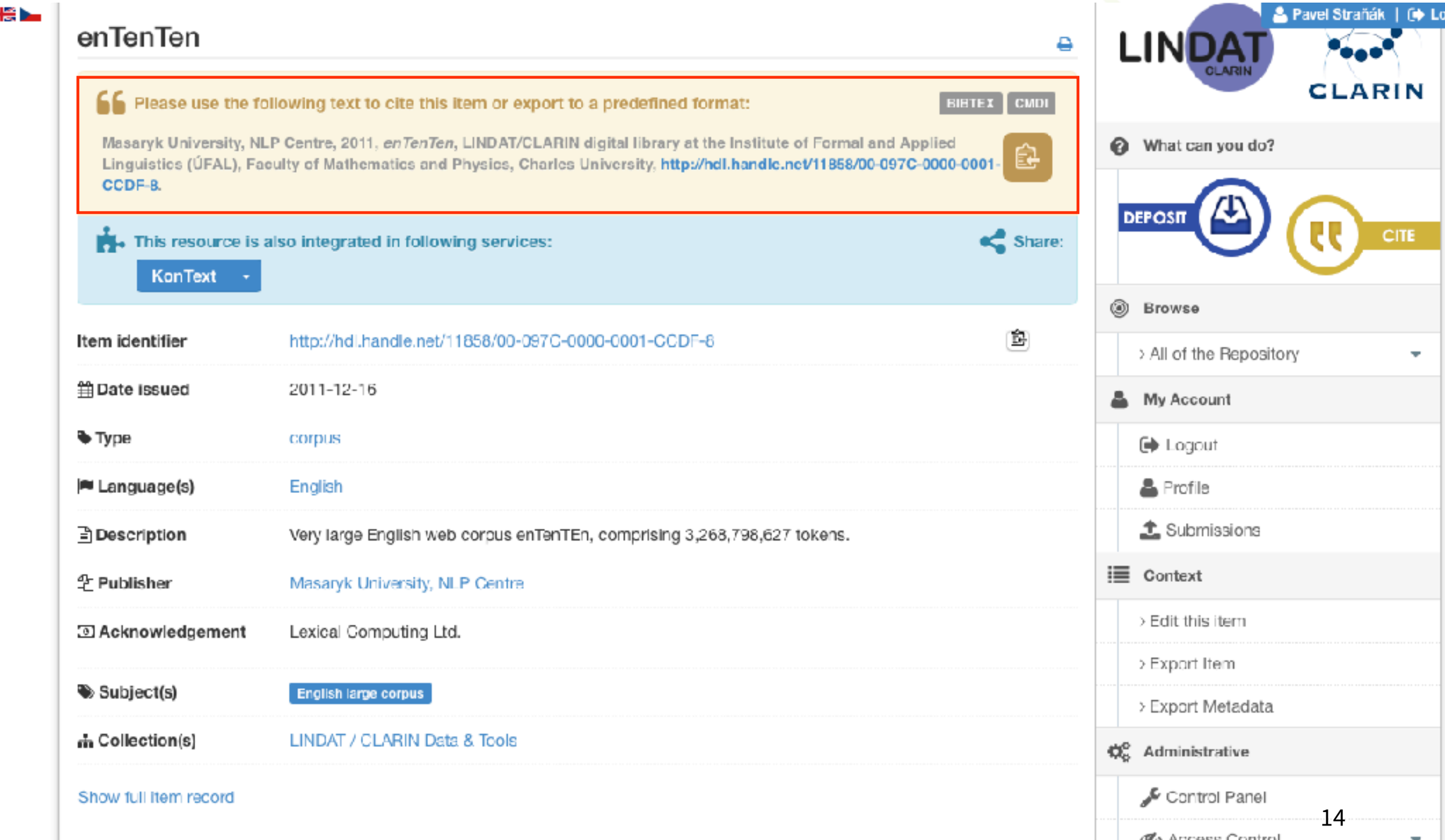
Discovery

Google



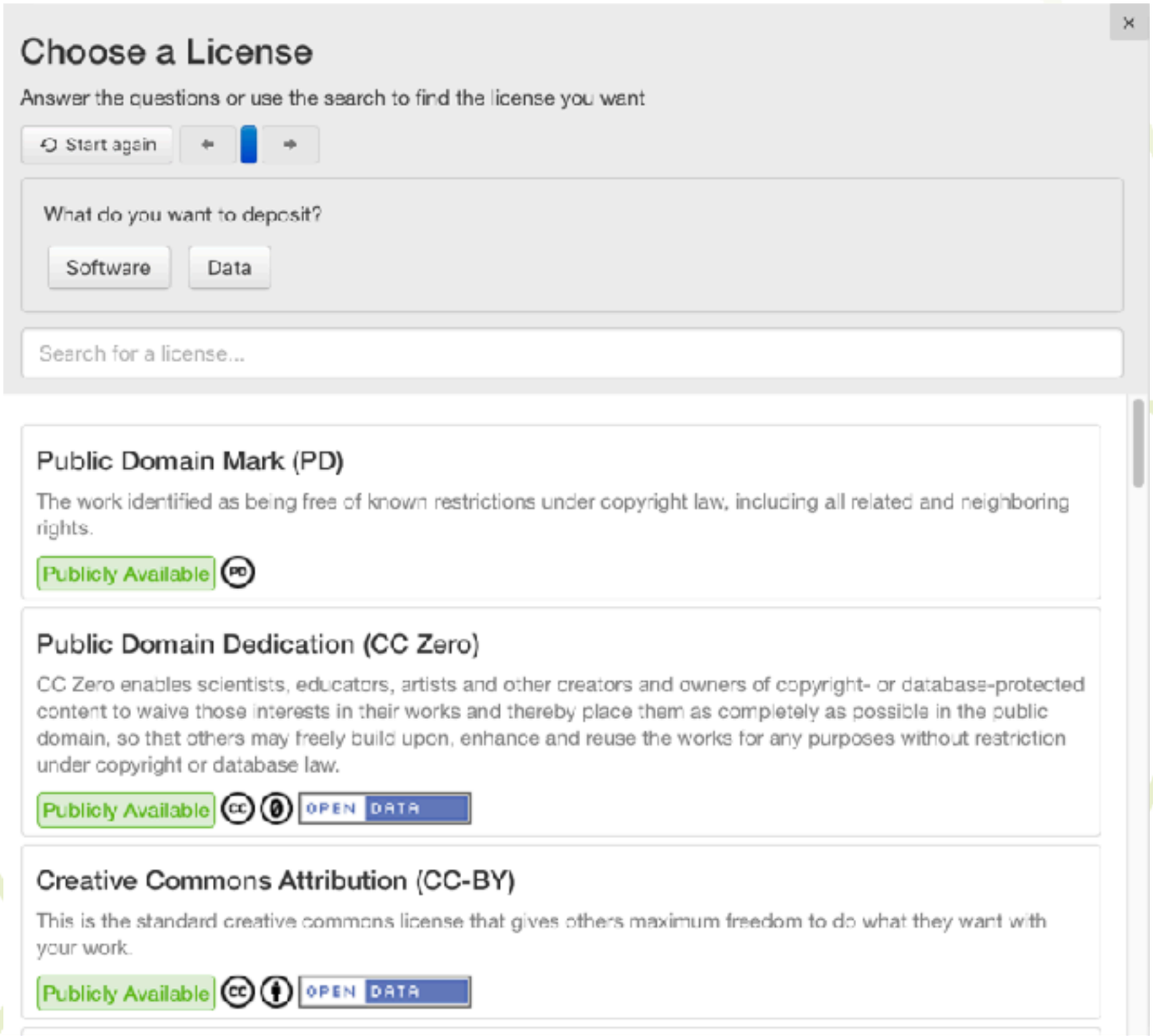
Direct Data Citations

Credit for Data



Licensing

As Open as Possible



Licensing

As Open as Possible (not more)

Publisher	Faculty of Arts, Institute of the Czech National Corpus, Charles University in Prague
Acknowledgement	Ministerstvo školství, mládeže a tělovýchovy Project code: LM2011023 Project name: Český národní korpus
Subject(s)	representative corpus · written language
Collection(s)	LINDAT / CLARIN Data & Tools

Show full item record

Files in this item

Download instructions for command line

This item is **Academic Use** and licensed under:
Czech National Corpus (Shuffled Corpus Data)

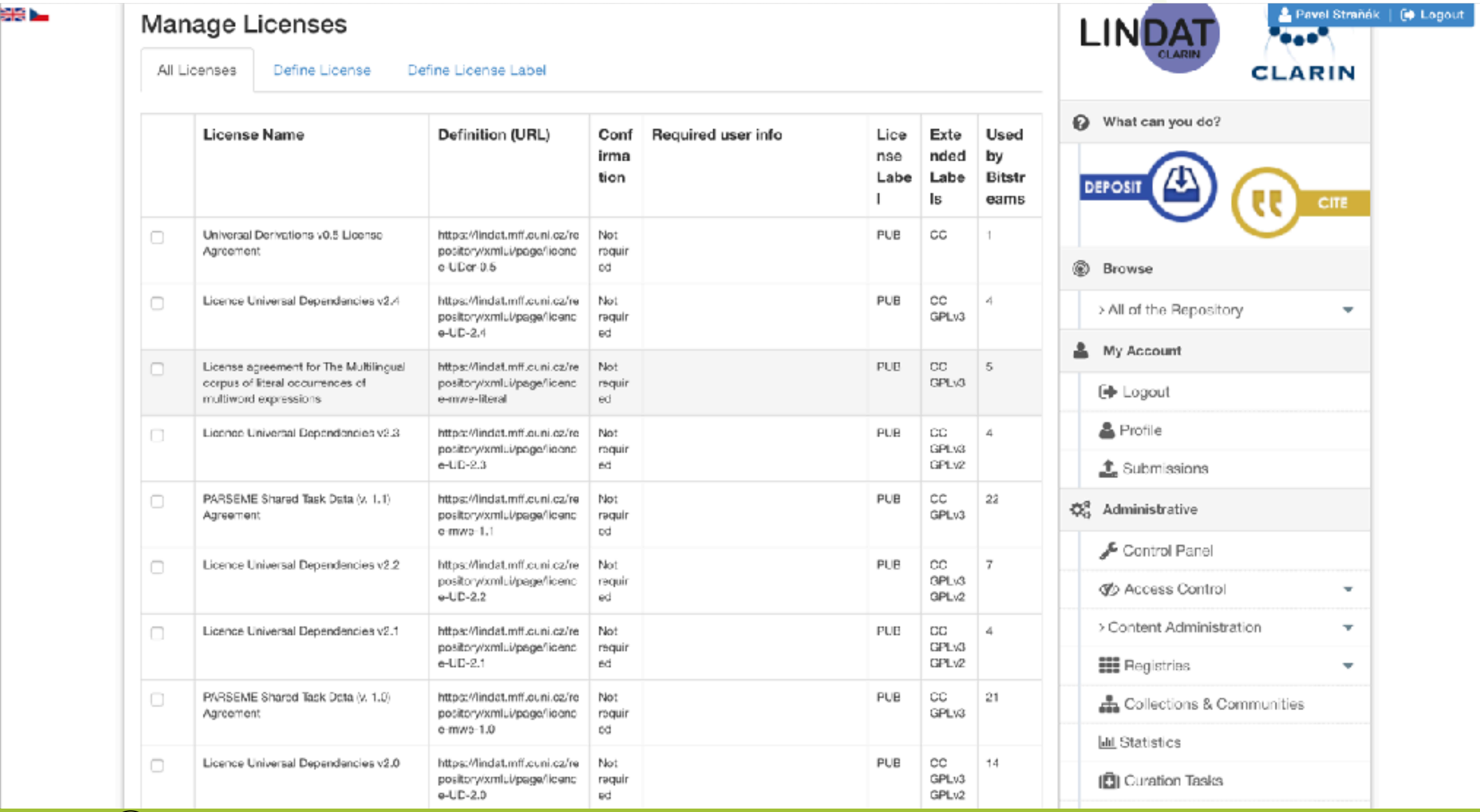
Name	syn2015.gz
Size	1.48 GB
Format	application/x-gzip
Description	corpus
MD5	e0242cc77e999794af6cfaf57f843c12

Download file

CLEAR RULES
CUSTOM LICENSES
LICENSE SIGNING

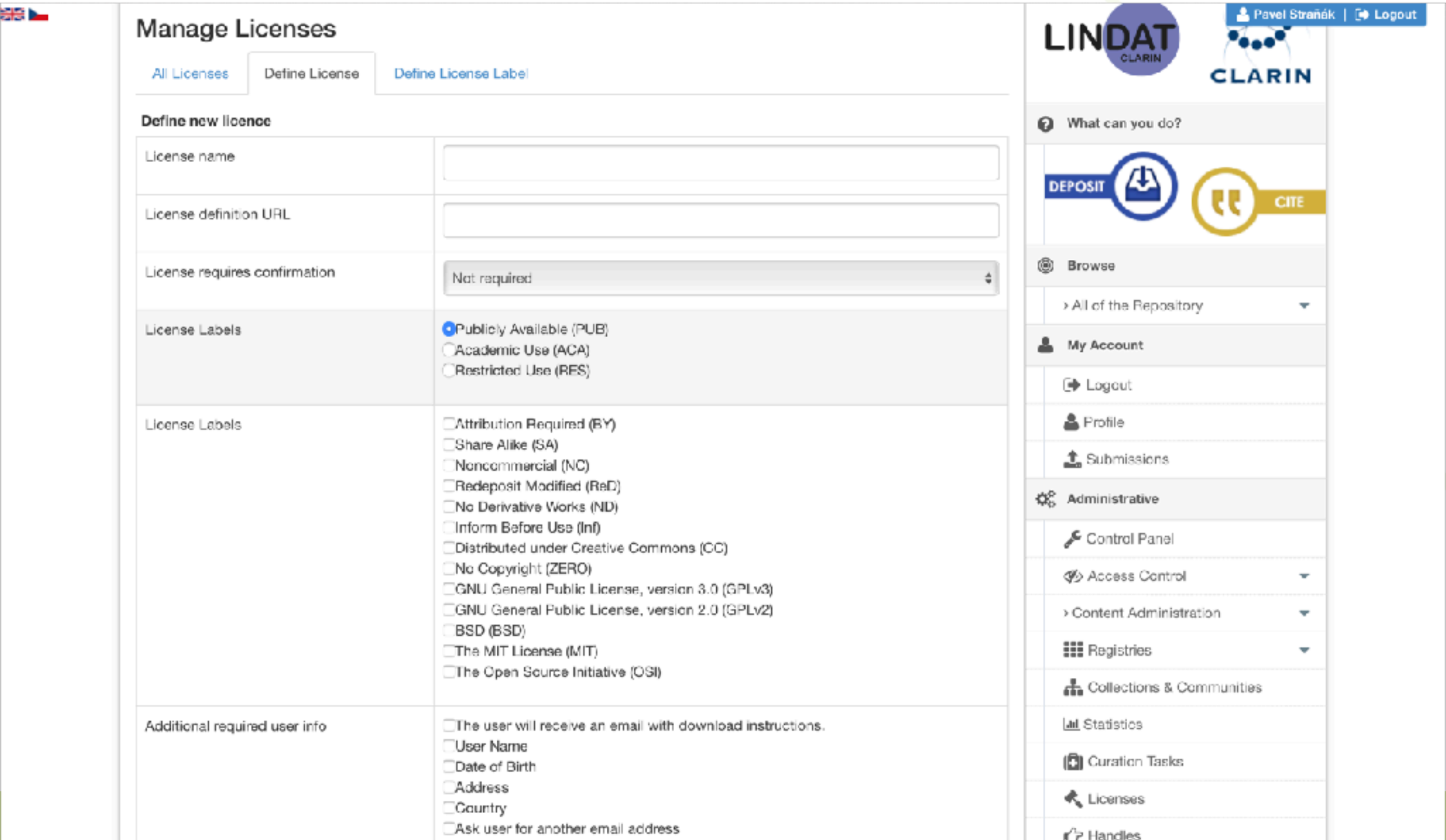
Licensing Framework:

Any License (Open Source / Open Data preferred)



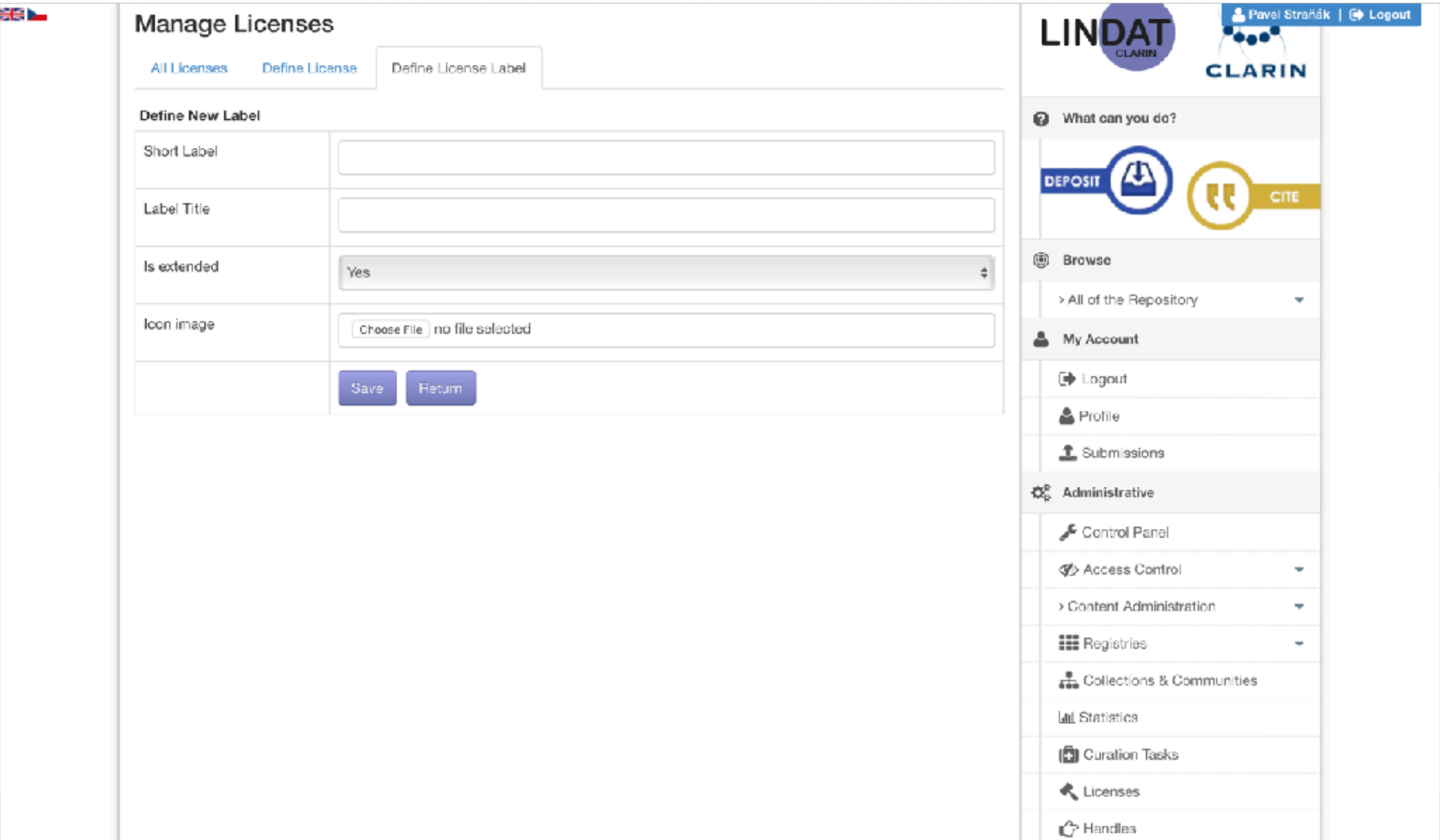
Licensing Framework

Defining a New License



Licensing Framework

Defining a licensing label / attribute



Versioning

prefer latest, preserve all

Project name: Internet jako jazykový korpus

Ministerstvo školství, mládeže a tělovýchovy České republiky

Project code: LN00A063

Project name: Centrum komputační lingvistiky

Ministerstvo školství, mládeže a tělovýchovy České republiky

Project code: MSM 0021620838

Project name: Moderní metody, struktury a systémy informatiky

Subject(s) MorphoDiTa | Czech | morphological analysis | morphological generation | PoS tagging

Collection(s) LINDAT / CLARIN Data & Tools

This item is replaced by a newer submission:
http://hdl.handle.net/11234/1-1836

Please refer to the submission above for the latest available data. If you nevertheless need the original data, please click here.

List all versions

Versioning

prefer latest, preserve all

Other versions

List all versions

- Czech Models (MorfFlex CZ 161115 + PDT 3.0) for MorphoDiTa 161115
- Czech Models (MorfFlex CZ 160310 + PDT 3.0) for MorphoDiTa 160310
- Czech Models (MorfFlex CZ + PDT) for MorphoDiTa

Show full item record

Files in this item

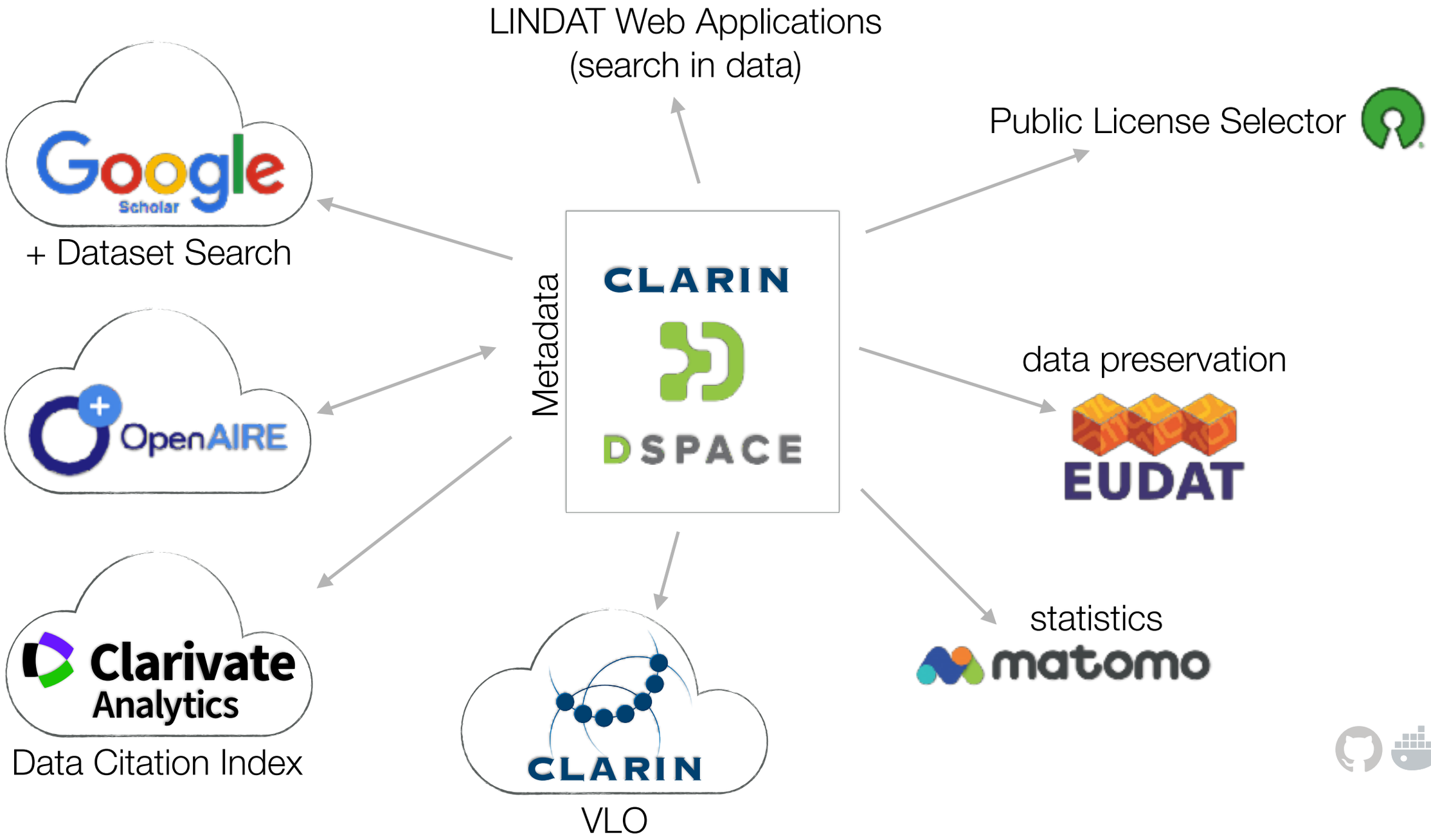
Download instructions for command line

This item is Publicly Available and licensed under:
Creative Commons - Attribution-NonCommercial-ShareAlike 4.0 International (CC BY-NC-SA 4.0)

Name	czech-morfflex-pdt-161115.zip
Size	69.18 MB
Format	application/zip
Description	Czech Models (MorfFlex CZ 161115 + PDT 3.0) for MorphoDiTa 161115
MD5	adde38cd363219759e19165b06baa4ce

Download file Preview

Repository Integrations



FAIR summary

- *Findable*: Google, Google Scholar, Data Citation Index, CLARIN VLO, OLAC… and the repository itself

- *Accessible*: records with data (even when restricted), complete licensing, Open Access (Public License Selector), login only when needed, CESNET, EUDAT

- *Interoperable*: common data formats, full documentation (enhanced metadata, documentation bitstreams)

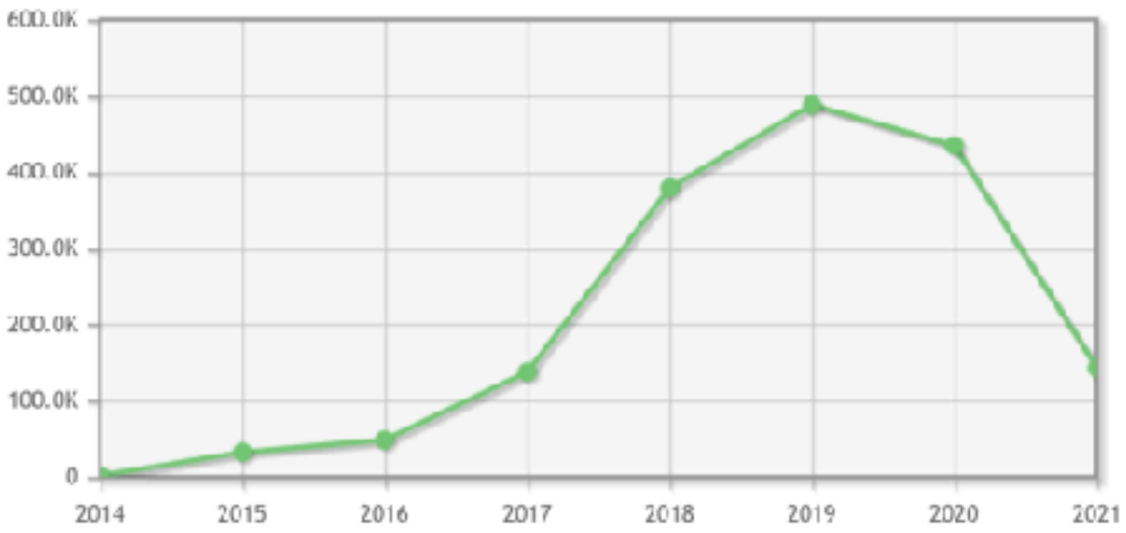
- *Reusable*: records with data, complete licensing, full versioning, direct data citations, maximal OA

Statistics

- 1123 records, 436 containing data:
    - 353 from LINDAT, 83 in LRT Inventory (worldwide open submissions)

The download statistics are based on the tracking request PIWIK received from the JAVA API directly from DSpace. The download counts also contain hits from bots.

Downloads Over Time 2014 - 2021



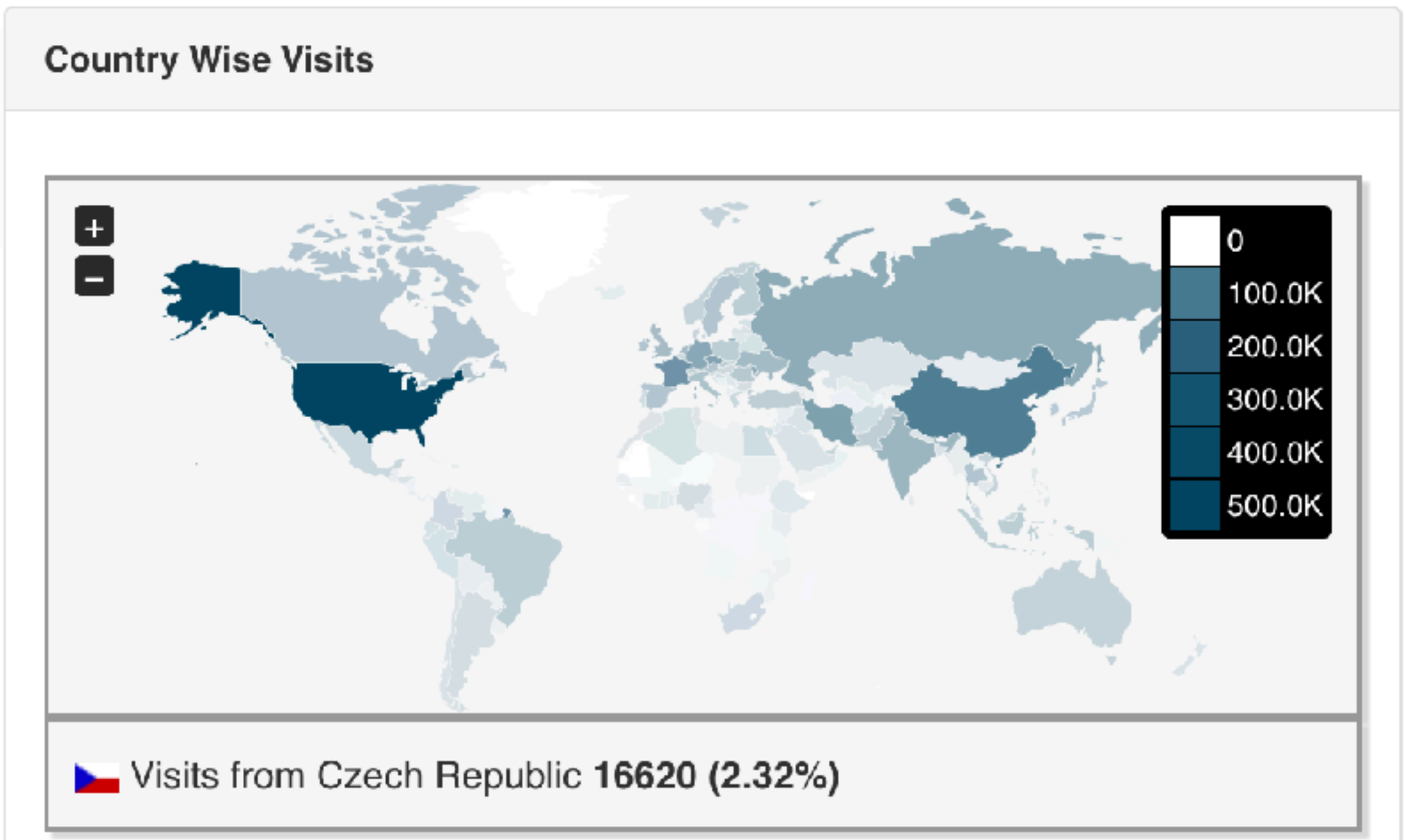
Click on a data item to summarize by year / month.

Important Metrics

Downloads	**1665033**	Unique Downloads	**1115926**
Visits	**717535**	Unique Visitors	**624774**

Hover on a metric name to show its definition.

Statistics


Country Wise Visits
0
100.0K
200.0K
300.0K
400.0K
500.0K
Visits from Czech Republic 16620 (2.32%)

Thank you!

https://lindat.cz/repository